Raymond Salzmann

11/1/2020

# Machine Learning Homework 2:

## Problem 1:

$$y_i \overset{ind}{\sim} N(x_i^T w, \lambda^{-1}) \quad w \sim N(0, \mathrm{diag}(\alpha_1, \dots, \alpha_d)^{-1})$$

$$\alpha_k = \mathrm{Gamma}(a_0, b_0) \qquad \lambda = \mathrm{Gamma}(e_0, f_0)$$

(a)

$$q(\lambda) \propto \exp\{\mathbb{E}_{-q(\lambda)}[\ln p(y \mid w, \alpha_{1:d}, \lambda, x)] + \ln q(\lambda)\}$$

$$\propto \exp\{\mathbb{E}_{q(w)}[\tfrac{N}{2}\ln(y) - \tfrac{\lambda}{2}(y - X^T w)^T (y - X^T w)] + (e_0 - 1)\ln(\lambda) - f_0 \lambda\}$$

$$\propto \exp\{(\tfrac{N}{2} + e_0 - 1)\ln\lambda - \tfrac{1}{2}\sum_{i=1}^{N}(\mathbb{E}_{q(w)}[(y_i - x_i^T w)^2] + 2f_0)\lambda$$

$$\propto \exp\{(\tfrac{N}{2} + e_0 - 1)\ln\lambda - (y^T y - 2y^T x^T \mathbb{E}_{q(w)}[w] + tr(\mathbb{E}_{q(w)}(ww^T) XX^T) + 2f_0)$$

$$q(\lambda) = \mathrm{Gamma}(e_0', f_0')$$

$$e_0' = e_0 + \tfrac{N}{2}$$

$$f_0' = \tfrac{1}{2}(y^T y - 2y^T x^T \mathbb{E}_{q(w)}[w] + tr(\mathbb{E}_{q(w)}[ww^T] XX^T) + f_0$$

$q(\alpha_k)$ where $k = 1{:}d$

$$q(\alpha_k) \propto \exp\{\mathbb{E}_{-q(\alpha_k)}[\ln p(w_k \mid \alpha_k)] + \ln q(\alpha_k)\}$$

$$\propto \exp\{\mathbb{E}_{q(w_k)}[\tfrac{1}{2}\ln(\alpha_k) - \tfrac{\alpha_k}{2}(w_k^2)] + (a_0 - 1)\ln\alpha_k - b_0\alpha_k\}$$

$$\propto \exp\left\{\left(\frac{1}{2}+a_0-1\right)\ln(\alpha_k) - \frac{1}{2}\left(\mathbb{E}_{q(w)}[w_k^2]+2b_0\right)\alpha_k\right\}$$

$$\boxed{\begin{aligned} q(\alpha_k) &= \text{Gamma}\left(a_0^{(k)'}, b_0^{(k)'}\right) \\ a_0^{(k)'} &= a_0 + \frac{1}{2} \\ b_0^{(k)'} &= \frac{1}{2}\left(\mathbb{E}_{q(w)}[w_k^2] + 2b_0\right) \end{aligned}}$$

$$q(w) \propto \exp\left\{\mathbb{E}_{-q(w)}\left[\ln p(y \mid \lambda, \alpha_{1:d}, w, x) + \ln q(w)\right]\right\}$$

$$\propto \exp\left\{\mathbb{E}_{q(\lambda)}\left[\frac{-\lambda}{2}(y - X^T w)^T (Y - X^T w)\right] + \mathbb{E}_{q(\alpha_{1:d})}\left[-\frac{\text{diag}(\alpha_1 \ldots \alpha_d)}{2} w^T w\right]\right\}$$

$$\propto \exp\left\{\frac{-1}{2} w^T\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right) w - 2 w^T \mathbb{E}_{q(\lambda)}[\lambda] X y\right\}$$

$$\propto \exp\left\{\frac{-1}{2}\left(w^T\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right) w - 2 w^T \left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \right.\right.\right.$$

$$\left. \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right) \mathbb{E}_{q(\lambda)}[\lambda]\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1} X y$$

$$+ \left[\mathbb{E}_{q(\lambda)}[\lambda]\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1} X y\right]^T$$

$$\cdot \left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right)$$

$$\cdot \left(\mathbb{E}_{q(\lambda)}\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1} X y\right)$$

Defining $A = \left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right)$ just for readability:

$$q(w) \propto \left\{\frac{-1}{2}\left(w^T A w - 2 w^T A\, \mathbb{E}_{q(\lambda)}[\lambda] A^{-1} X y + \left(\mathbb{E}_{q(\lambda)}[\lambda] A^{-1} X y\right)^T A \left(\mathbb{E}_{q(\lambda)}[\lambda] A^{-1} X y\right)\right)\right\}$$

$$q(w) = \text{Normal}\left(\mathbb{E}_{q(\lambda)}[\lambda]\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1} X y,\right.$$

$$\left.\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\text{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1}\right)$$

$$q(\omega) = N\left(E_{q(\lambda)}[\lambda] A^{-1} X Y, A^{-1}\right)$$

$$\mu' = E_{q(\lambda)}[\lambda] A^{-1} X y$$

$$\Sigma' = A^{-1} = \left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\mathrm{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1}$$

Defining expectations that have been used and are needed to update the various q distributions:

(1) $\mathbb{E}_{q(\alpha_{1:d})}[\mathrm{diag}(\alpha_1, \ldots, \alpha_d)] = \mathrm{diag}\left(a_0^{(1)'}/b_0^{(1)'}, \ldots, a_0^{(d)'}/b_0^{(d)'}\right)$

(2) $\mathbb{E}_{q(\lambda)}[\lambda] = e_0'/f_0'$

(3) $\mathbb{E}_{q(w_k)}[w_k^2] = \left(\mathbb{E}_{q(\omega)}[\omega]\right)_k^2 + \left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\mathrm{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1}$

(4) $E_{q(\omega)}[\omega \omega^T] = \left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\mathrm{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1} + E_{q(\omega)}[\omega] E_{q(\omega)}[\omega^T]$

(5) $E_{q(\omega)}[\omega] = E_{q(\lambda)}[\lambda]\left(\mathbb{E}_{q(\lambda)}[\lambda] X X^T + \mathbb{E}_{q(\alpha_{1:d})}[\mathrm{diag}(\alpha_1, \ldots \alpha_d)]\right)^{-1} X y$

(b)

Step 1: Initialize prior parameters $a_0', b_0', e_0', f_0', \mu_0'$ and $\Sigma_0'$

Step 2: For iteration $t = 1, \dots, T$

(i) Update $q(\lambda)$:

$$- e_t' = \frac{N}{2} + e_0$$

$$- f_t' = \frac{1}{2}\left(y^T y - 2 y^T x^T (\mu_{t-1}) + tr\left((\Sigma_{t-1} + \mathbb{E}_{q(w)}[w_t]\,\mathbb{E}_{q(w)}[w]^T\right)\right)$$

(ii) Update $q(\alpha_k)$:

$$- a_t^{(k)'} = \frac{1}{2} + a_0$$

$$- b_t^{(k)'} = \frac{1}{2}\left(\mathbb{E}_{q(w)}[w_t]\right)_k^2 + (\Sigma_{t-1})_{kk}$$

(iii) Update $q(w)$ : $A = \left(\mathbb{E}_{q(\lambda)}[\lambda] XX^T + \mathbb{E}_{q(\alpha_{1:d})}[\mathrm{diag}(\alpha_1, \dots \alpha_d)]\right)$

$$- A_t' = \mathbb{E}_{q(\lambda)}[\lambda_t] XX^T + \mathbb{E}_{q(\alpha_{1:d})}[\mathrm{diag}(\alpha_1, \dots \alpha_d)_t]$$

$$- \Sigma_t' = A_t^{-1}$$

$$- \mu_t' = \mathbb{E}_{q(\lambda)}[\lambda_t]\, A_t^{-1} XY$$

(iv) Calculate the variational objective function ; ELBO

$$\mathcal{L}(a_t', b_t', e_t', f_t', \mu_t', \Sigma_t')$$

(c) $\mathcal{L}((a_t^{(1)'}, b_t^{(1)'}), \ldots, (a_t^{(d)'}, b_t^{(d)'}), e_t', f_t', \mu_t', \Sigma_t')$

$$= \int_w \int_\lambda \int_{\alpha_{1:d}} q(w) q(\lambda) \prod_{k=1}^{d} q(\alpha_k) \ln \frac{p(w, \alpha_{1:d}, \lambda \mid Y, X)}{q(w) q(\lambda) \prod_{k=1}^{d} q(\alpha_k)} d\alpha_{1:d}\, d\lambda\, dw$$

$$= \overset{(1)}{\int_w \int_{\alpha_{1:d}} q(w) \ln p(w) d\alpha_{1:d} dw} + \overset{(2)}{\int_\lambda q(\lambda) \ln p(\lambda) d\lambda} + \overset{(3)}{\int_{\alpha_{1:d}} \prod_{k=1}^{d} q(\alpha_k) \ln p(\alpha_k) d\alpha_{1:d}}$$

$$+ \overset{(4)}{\iint_{w\,\lambda} q(w) q(\lambda) \ln p(y \mid w, \lambda, X) d\lambda dw} - \overset{(5)}{\int_w q(w) \ln q(w) dw}$$

$$- \overset{(6)}{\int_\lambda q(\lambda) \ln q(\lambda) d\lambda} - \overset{(7)}{\int_{\alpha_{1:d}} \prod_{k=1}^{d} q(\alpha_k) \ln q(\alpha_k) d\alpha_{1:d}}$$

Eqn (1): $q(w) \ln p(w)$

$$\iint_{w\,\alpha_{1:d}} q(w) \ln p(w) dw = \frac{-d}{2} \ln(2\pi) + \frac{1}{2} \sum_{k=1}^{d} \mathbb{E}_{\alpha_{1:d}}[\ln(\alpha_k)] - \frac{1}{2} \mathbb{E}_{q(w), q(\alpha_1, \alpha_d)}[w_t^T \operatorname{diag}(\alpha_1, .. \alpha_d) w_t]$$

$$= \frac{-d}{2} \ln(2\pi) + \frac{1}{2} \sum_{k=1}^{d} (\psi(a_t^{(k)'}) - \ln(b_t^{(k)'})) - \frac{1}{2} \operatorname{tr}\left((\mu_t' \mu_t'^T + \Sigma_t') \mathbb{E}_{q(\alpha_{1:d})}[\operatorname{diag}(\alpha_1, .. \alpha_d)]\right)$$

Eqn (2): $q(\lambda) \ln p(\lambda)$

$$\int_\lambda q(\lambda) \ln p(\lambda) d\lambda = e_0 \ln(f_0) - \ln(\Gamma(e_0)) + (e_0 - 1) \mathbb{E}_{q(\lambda)}[\ln \lambda] - f_0 \mathbb{E}_{q(\lambda)}[\lambda]$$

$$= e_0 \ln(f_0) - \ln(\Gamma(e_0)) + (e_0 - 1)(\psi(e_t') - \ln(f_t')) - f_0 \frac{e_t'}{f_t'}$$

Eqn (3): $\prod_{k=1}^{d} q(\alpha_k) \ln p(\alpha_k)\, d_{\alpha_{1:d}}$

$$\int_{\alpha_{1:d}} \prod_{k=1}^{d} q(\alpha_k) \ln(p(\alpha_k))\, d_{\alpha_{1:d}} = \int_{\alpha_{1:d}} \sum_{k=1}^{d} (\alpha_k) \ln(p(\alpha_k))\, d\alpha_k$$

$$= \sum_{k=1}^{d} \int_{\alpha_k} (\alpha_k) \ln(p(\alpha_k))\, d\alpha_k$$

$$= \sum_{k=1}^{d} a_0 \ln(b_0) - \ln(\Gamma(a_0)) + (a_0 - 1)(\psi(a_t^{(k)\prime}) - \ln(b_t^{(k)\prime})) - b_0 \frac{a_t^{(k)\prime}}{b_t^{(k)\prime}}$$

Eqn (4): $q(w) \ln(w)\, dw$

$$\int_w q(w) \ln(w)\, dw = \frac{1}{2} \ln((2\pi e)^n |\Sigma_t|)$$

Eqn (5): $q(\lambda) \ln(\lambda)\, d\lambda$

$$\int_\lambda q(\lambda) \ln(\lambda)\, d\lambda = e_t' \ln(f_t') - \ln(\Gamma(e_t')) + (e_t' - 1)(\psi(e_t') - \ln(f_t')) - f_t' \frac{e_t'}{f_t'}$$

$$= \ln(f_t') - \ln(\Gamma(e_t')) + (e_t' - 1)\psi(e_t') - e_t'$$

Eqn (6): $\prod_{k=1}^{d} q(\alpha_k) \ln(q(\alpha_k))$

$$\int_{\alpha_{1:d}} \prod_{k=1}^{d} q(\alpha_k) \ln q(\alpha_k)\, d_{\alpha_{1:d}} = \sum_{k=1}^{d} \ln(b_t^{(k)\prime}) - \ln(\Gamma(a_t^{(k)\prime})) + (a_t^{(k)\prime} - 1)\psi(a_t^{(k)\prime}) - a_t^{(k)\prime}$$

Eqn (7): $q(\omega)q(\lambda)\ln(p(y|\omega, X, \lambda))$

$$\int_\omega\int_\lambda q(\omega)q(\lambda)\ln(p(y|\omega,\lambda,X))\,d\lambda\,d\omega$$

$$= \frac{-N}{2}\ln(2\pi) + \frac{N}{2}\left(\mathbb{E}_{q(\lambda)}[\lambda]\right) - \frac{\mathbb{E}_{q(\lambda)}[\lambda]}{2}\mathbb{E}_{q(\omega)}[\omega]$$

$$\cdot\left[(y-X^T\omega)^T(y-X^T\omega)\right]$$

$$= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\left(\psi(e_t') - \ln(f_t')\right) - \frac{e_t'}{2f_t'}\left[y^Ty - 2y^TX^T\mu_t' - tr\left((\Sigma_t' + \mu_t'\mu_t'^T)\right.\right.$$

$$\left.\left.\cdot XX^T\right)\right]$$

$\therefore$ In order to get the entirety of the variational objective function, we must combine the final forms of the 7 parts from the original equation which we have broken down above.

Thus, we have: For ease of readability, the different parts will be different colors:

$$\mathcal{L}\left((a_t^{(1)'}, b_t^{(1)'}) \ldots (a_t^{(d)'}, b_t^{(d)'}), e_t', f_t', \mu_t', \Sigma_t'\right) =$$

$$= \frac{-d}{2}\ln(2\pi) + \frac{1}{2}\sum_{k=1}^{d}\left(\psi(a_t^{(k)'}) - \ln(b_t^{(k)'})\right) - \frac{1}{2}tr\left((\mu_t'\mu_t'^T + \Sigma_t')\mathbb{E}_{q(\alpha_{1:d})}[diag(\alpha_1, \ldots \alpha_d)]\right)$$

$$+ e_0\ln(f_0) - \ln(\Gamma(e_0)) + (e_0 - 1)(\psi(e_t') - \ln(f_t')) - f_0\frac{e_t'}{f_t'}$$

$$+ \sum_{k=1}^{d} a_0\ln(b_0) - \ln(\Gamma(a_0)) + (a_0 - 1)(\psi(a_t^{(k)'}) - \ln(b_t^{(k)'})) - b_0\frac{a_t^{(k)'}}{b_t^{(k)'}}$$

$$+ \frac{1}{2}\ln\left((2\pi e)^n |\Sigma_t|\right)$$

$$- \left\{ \ln(f_t') - \ln\left(\Gamma(e_t')\right) + (e_t' - 1)\psi(e_t') - e_t' \right\}$$

$$- \left\{ \sum_{k=1}^{d} \ln\left(b_t^{(k)'}\right) - \ln\left(\Gamma\left(a_t^{(k)'}\right)\right) + \left(a_t^{(k)'} - 1\right)\psi\left(a_t^{(k)'}\right) - a_t^{(k)'} \right\}$$

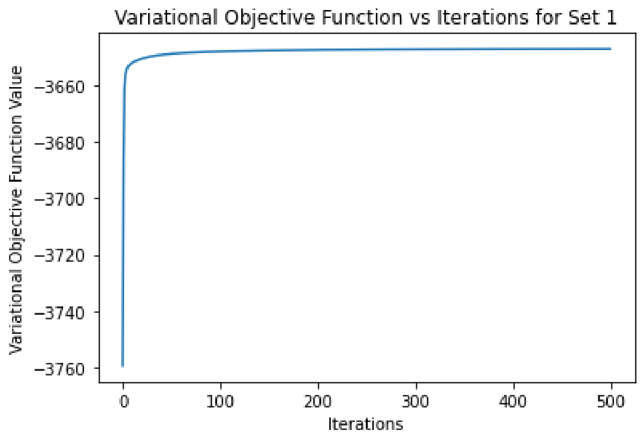
$$- \left\{ -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\left(\psi(e_t') - \ln(f_t')\right) - \frac{e_t'}{2f_t'}\left[ Y^T y - 2y^T X^T \mu_t' - tr\left((\Sigma_t' + \mu_t' \right.\right.\right.$$

$$\left.\left.\left. \circ XX^T\right)\right] \right\}$$

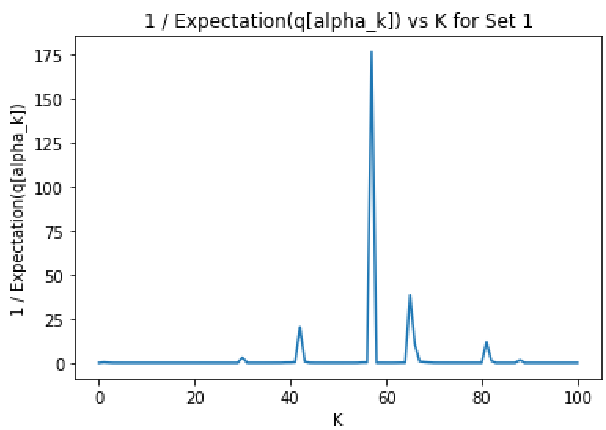
Thus, with the above ELBO, we can check the convergence of our pseudo-code algorithm in part b above.

**Question 2:**

Part A (Set 1)



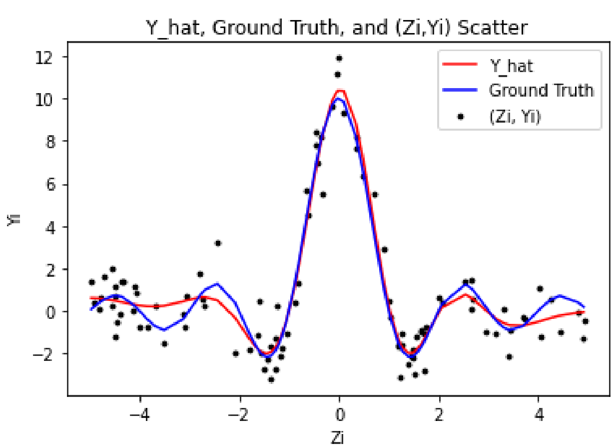

Part B (Set 1)



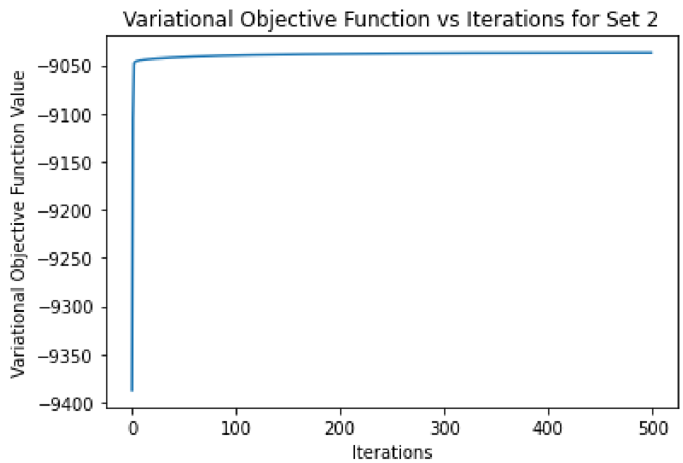

Part C (Set 1)

```
1/E_q[lambda] for Set 1 =  [[1.01909983]]
```

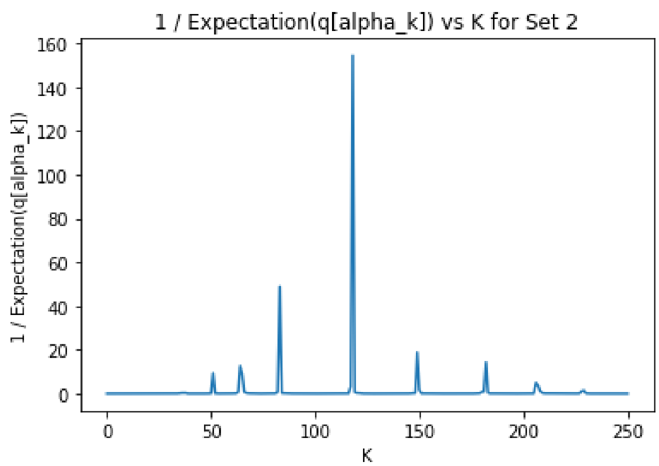
Part D (Set 1)



Part A (Set 2)



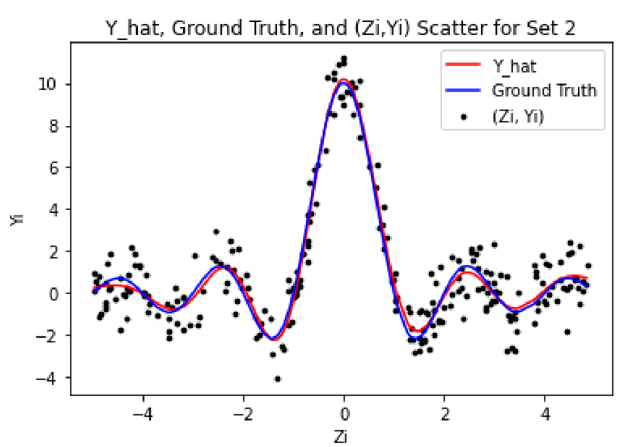

Part B (Set 2)



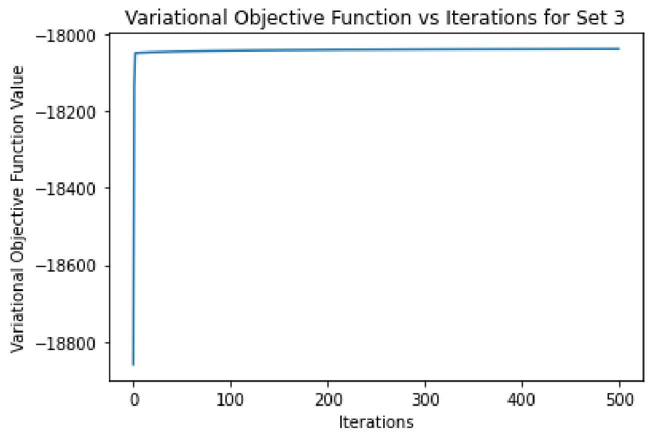

Part C (Set 2)

```
1/E_q[lambda] for Set 2 =  [[0.90035679]]
```

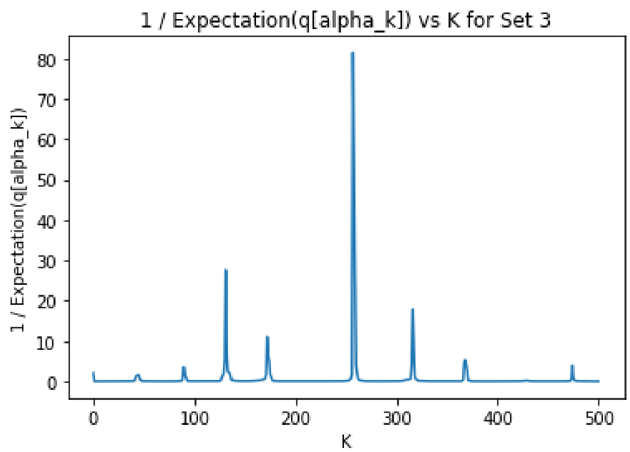
Part D (Set 2)



Part A (Set 3)




Part B (Set 3)




Part C (Set 3)

```
1/E_q[lambda] for Set 3 =  [[0.972746]]
```

Part D (Set 3)